Stripping flow cytometry: how many detectors do we need for bacterial identification?

Peter Rubbens[1*], Ruben Props[2], Cristina Garcia-Timermans[2], Nico Boon[2], Willem Waegeman[1]

[1]KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University

[2]Center for Microbial Technology and Ecology (CMET), Ghent University

Ghent, Belgium

[*]Corresponding author

Tel.: (+32) 9 264.60.18

Fax: (+32) 9 264.62.20

E-mail: Peter.Rubbens@UGent.be.

Keywords

Automated identification of bacterial populations; Bacterial communities; Detector elimination; Flow cytometry; Microbiology; Single-cell analysis; Synthetic microbiology; Variable selection.

Abstract

Multicolor approaches are challenging for microbial flow cytometry; as flow cytometers are mainly developed for biomedical applications, modern instruments contain more detectors than needed. Some of these additional fluorescence detectors measure biological information due to spectral overlap, yet the extent to which this information is relevant for the identification of bacterial populations is ambiguous. In this paper we characterize the usefulness of these additional detectors. We propose a data-driven detector selection method to select the smallest subset of detectors that will optimally discriminate between bacterial populations. Using a detector elimination strategy, we show that one or more detectors can be removed without loss of resolving power. A number of additional detectors are included in the final subset, which help to improve the identification of bacterial populations. Experimental data were retrieved from two types of modern cytometers with different configurations. The method reveals a clear ordering of detector importances, which depends on the instrument from which the data were retrieved. In addition we were able to pinpoint unexpected behavior of SYBR Green I in the red spectrum. As the field of microbial flow cytometry is maturing, these results motivate the construction of a different kind of cytometric instruments for microbiologists, for which the number of detectors is reduced, but tailored towards the characteristics of microbial experiments.

Introduction

Flow cytometry (FCM) is a well-established method for the analysis of microbial communities. Originally used as a tool to assess bacterial heterogeneity and viability [1], FCM has shown its significance for both environmental applications and industrial setups [2,3]. In recent literature, more and more emphasis is being placed on the study of synthetic microbial communities [4,5]. Typically, these communities contain a lower amount of bacterial species. They exhibit key features of their natural counterpart community, but are created and studied in a highly-controlled environment. Therefore, they can serve as a proxy between microbial theories on the one hand and real natural communities on the other hand [6,7]. Recently we have been able to use so-called *in silico* communities to retrieve the composition of low-complexity synthetic communities using FCM in combination with a machine learning based approach [8]. This approach makes use of an *in silico* data-aggregation step, which allows us to benefit from the availability of species labels and therefore enables the use of supervised machine learning methods. As *in silico* communities have proven to be a valid stand-in for synthetic microbial communities, they can be further exploited by adopting a data-driven approach in function of research questions in the field of microbial FCM.

Microbial FCM suffers to a greater extent from technical and biological limitations as compared to biomedical applications [9]. Staining bacteria is subject to a complex interplay between dye chemistry, target organisms and staining conditions. For microbiological applications, the diversity of bacterial species is challenging, as even closely related organisms are known to possess varying physiological characteristics [10]. Therefore it is difficult to analyze bacteria in a standardized way [11]. Additional complications arise due to cell sizes, which are much smaller compared to mammalian cells [12–14]. This is why most microbial flow cytometry experiments make use of one or two stains. One expects therefore that microbial FCM exper-

iments result in three or four parametric data sets at best, containing forward and side scatter information, combined with one or two fluorescence signals. Yet, driven by human research, modern flow cytometers are equipped with more detectors [15], which is why more information than often necessary is measured in current practices. This means that when applying microbial FCM, some additional not-targeted fluorescence detectors measure leakage coming from the targeted channel due to spectral overlap. This is often defined as *cross-talk* or *spillover* between detectors. As this information is often neglected based on a theoretical point of view, most researchers are interested in compensating for this effect in multicolor experiments [16–18]. Some microbial procedures make use of a secondary detector for denoising purposes [19–21], but little research has been devoted to an actual characterization of the relevance of these additional detectors.

The objective of this paper is to quantify the usefulness of all detectors on modern flow cytometers. We propose a machine learning based detector elimination strategy which allows us to objectively decide which detectors to retain and to quantify their importance in function of bacterial identification. Our method initially considers all available detectors. Next, the detector that has the lowest resolving power is incrementally removed. In an artificial way, flow cytometric data is stripped sequentially from its least effective detectors. Our detector elimination strategy was applied on data derived from two types of cytometers with different specifications which analyzed biological replicates stained with SYBR Green I.

Materials & Methods

Dataset description

FCM data of 20 individual bacterial cultures stained and analyzed with SYBR Green I (Invitrogen), as previously described in [8], were retrieved from FlowRepository (ID: FR-FCM-ZZSH). In brief, samples were diluted to approximate cell densities of 10^6 cells mL^{-1} in 0.22 µm filtered PBS (6.8 $gL^{-1}KH_2PO_4$, 8.8 gL^{-1} K_2HPO_4 and 8.5 gL^{-1} NaCl) and stained with a final concentration of 1% (v/v) nucleic acid stain SYBR Green I (100x concentrate in 0.22 µm filtered dimethyl sulfoxide). Samples were incubated for 20 minutes in the dark at 37 °C and immediately analyzed by means of an autoloader. All cultures were sampled after 24h of incubation. The growth curves of each culture indicate that most cultures (n = 17) were in early-to-mid stationary phase, while a few (n = 3) were still in the exponential or linear growth phase at the time of sampling (SI Fig. 5).

The samples were analyzed by an Accuri C6 flow cytometer (BD Biosciences) at 66 μL/min and FL1-H threshold of 500. Prior to measurement, the performance of the Accuri C6 was evaluated by analyzing eight peak rainbow particles (Spherotech, Lake Forest, IL, USA). The performance check was passed if each bead population was located at its fixed position and displayed a coefficient of variation on its specific fluorescence channels of $< 5\%$. Samples were analyzed in fixed volume mode (50 μL per sample) after 20 minutes incubation in the dark to ensure the reproducibility of the staining protocol. Biological replicates were analyzed on a FACSVerse flow cytometer at 60 μL/min for a maximum of 1 minute (BD Biosciences) (FlowRepository ID: FR-FCM-ZY6M); see Tab. 1 for an overview of the detector setup for both instruments, along with an estimation of the theoretical filter leakage due to spectral overlap for SYBR Green I. The performance of the FACSVerse was verified by the FACSuiteTM software performance quality check using CS&T research beads (BD Biosciences). The quality check

compares the flow cytometry data of CS&T research beads with the previous recorded bead data. Significant deviations from the bead parameter values at the detector and laser parameters predefined for this specific experiment would cause the quality check to fail. For a full technical overview we refer to the manuals [22, 23].

Instrumental and (in)organic noise were removed using a reproducible digital gating strategy in the $\mathrm{arcsinh}(x)$ transformed FL1 – FL3 (or FITC – PerCP-Cy5.5 equivalent) bivariate space [19,20]. This filtering strategy was verified by negative controls (non-stained samples) and kept fixed for all samples of the same individual culture. Results of the denoising can be found in the supplementary information (SI Fig. 6) for the FACSVerse data; for the Accuri C6 they can be consulted in [8]. An additional stringent three-step data-driven denoising was applied on the filtered data in order to remove cells for which there was erroneous parameter acquisition using the automated flowAI package (v1.4.4., default settings, target channel = FL1 or FITC, changepoint detection penalty for Accuri = 150, for FACSVerse = 200) [24]. In short, flowAI removes anomalous events in function of three stability criteria: (1) the flow rate, expressed by the number of cells per unit of time, (2) signal acquisition, defined by a stable average fluorescence intensity per unit of time and (3) the dynamic range, removing margin events that lie higher than the dynamic range of a flow cytometer and that are therefore accumulated in the last channel of the dynamic range.

In silico communities

We created *in silico* communities to employ our detector elimination strategy. This means that communities were created artificially by aggregating data coming from bacterial cultures, which were measured individually. These *in silico* communities have proven to be a valid representation of synthetic microbial communities [8]. Our *in silico* approach benefits from two advantages: we are able to evaluate our strategy on a great amount of possible communities, an

amount which is much larger than is feasible in the lab. This enables us to draw more general conclusions. Second, we are able to exploit the labels of bacterial single cells, which enables to use supervised machine learning methods to identify single cells. This allows us to capture relations between variables, in this case detectors, which unsupervised statistical models are not able to.

In silico communities were created for various species richness S, i.e., the number of bacterial populations present in a community. For $S = 2$ and $S = 18$ all possible community compositions at the species level were evaluated, which is 190. Communities were also created for $S = 6, 10$ or 14, for which 190 different bacterial compositions were drawn at random. Per replicate we sampled 5,000 cells, adding 2,500 cells to a training and test set respectively. As we have two replicates per individual culture at our disposal, the number of cells N in a training and test set equals 5,000 cells times the number of bacterial populations present. The same community compositions were evaluated for the two types of datasets.

Random Forest classifier

We used a Random Forest classifier in order to classify bacterial single cells [26]. The Random Forest algorithm is an ensemble method, which uses a decision tree as base classifier. It makes use of two kinds of randomization in order to reduce the variance of the predicted output. First, it fits a fully grown decision tree to $n = 200$ bootstrap samples. Second, a decision tree only gets to choose from a random subset of a total of K variables at every split. Our choice for the algorithm is motivated by the fact that Random Forests have shown to be a reliable method to retrieve the community composition of a synthetic community [8]. It belongs to the top-performing 'off-the-shelf' classifiers [27] and is an established method in the field of computational biology [28]. Moreover, it inherits a number of favorable properties of decision trees, such as the fact that decision trees are insensitive to transformations of the data and that

it is able to handle multiclass datasets in a natural way. Usually, the Random Forest classifier does not suffer from correlated variables. There was no need to tune its hyperparameter K (SI Fig. 7), which is why we used the preset $\sqrt{K}$, along with default settings. Therefore computational costs remain low while achieving a high performance. The identification of bacterial populations was evaluated in terms of the accuracy, which expresses the fraction of cells that were classified correctly. The machine learning library *scikit-learn* (v0.18) was used to perform the analysis [29].

Detector elimination strategy

The goal of this paper is to investigate how many detectors can be eliminated while retaining an optimal performance concerning the identification of a bacterial community. In order to be able to incorporate higher-order interactions between detectors, we implemented a wrapper method, using a backward stepwise elimination strategy [30]. This means that an analysis was started with the incorporation of all detectors. Next, the detector which gave the smallest drop in bacterial identification accuracy was removed from the dataset in an incremental fashion, until there was one detector left. This approach implies that all parameters from a single detector were used, i.e. both the area, height and for the FACSVerse the width parameter. The longer a detector is retained in the analysis, the more important it is considered to be. A formal scheme of the elimination strategy can be found in Algorithm 1.

Results

Mutual variable correlations

Staining bacteria with SYBR Green I targets the FL1- and FITC-detector for the Accuri C6 and FACSVerse respectively. Based on the theoretical estimated filter leakage, one expects one (Accuri C6) or five (FACSVerse) detectors to measure additional information due to cross-talk (Tab. 1). Mutual variable dependencies, in terms of the pearson correlation ρ, were calculated in order to quantify the actual amount of additional information that is measured by both cytometers. In this way we were able to assess to what extent secondary signals were correlated with the target detector based on experimental values. This was done for all samples ($n = 40$ for each instrument) and averaged using a Fisher transformation (Fig. 1).

This preliminary analysis illustrates that actual variable dependencies only partially comply with dependencies based on theoretically estimated cross-talk. Inspecting the Accuri C6 cytometer, we see that all secondary fluorescence detectors were significantly correlated to the target detector (i.e., significantly correlated with at least one channel area, height or width of the target fluorescence detector, $\rho > 0.41$, $P < 0.01$, using a one sided Z-test), especially the FL2 and FL3 detectors. This was unanticipated, as only FL2 was expected to measure information

Algorithm 1: Detector elimination scheme

input : training set, test set, list of detectors $D = \{d_1, ..., d_D\}$;
output: ranking of detectors R;
calculate performance RandomForestClassifier(train, test, D);
while $|D| > 0$ **do**
  for $d \in D$ **do**
    $D' \leftarrow$ remove d from D;
    calculate performance RandomForestClassifier(train, test, D');
  $D \leftarrow$ remove detector d_l with lowest resolving power from D;
  update R;

due to spectral overlap. For the FACSVerse cytometer, four out of five expected fluorescence detectors detectors showed significant correlations to the target detector ($\rho > 0.41$, $P < 0.01$, using a one sided Z-test), the exception being the V450-detector. In general, we note that experimental cross-talk did not match with what was expected from theoretical estimations for SYBR Green I.

Single detector identification performance

First, bacterial populations were identified feeding information coming from a single detector only to the Random Forest algorithm (Fig. 2, SI Fig. 8). Doing so allows one to compare detectors directly and to fully assess the resolving power a single detector is able to capture.

Secondary fluorescence detectors that were significantly correlated to the target detector were able to identify bacterial populations better than random guessing ($\rho > 0.41$, $P < 0.01$, using a one sided Z-test). The secondary detector which is closest to the target detector was able to identify bacterial single cells with an equivalent resolving power. Although a higher correlation generally gave rise to a higher identification capacity, this ranking was not strict (the exception being the V500-detector). We conclude that secondary detectors that captured cross-talk can be used for the identification of bacterial cells.

Both forward and side scatter detectors of the FACSVerse cytometer are able to distinguish bacterial single cells with equivalent accuracy as the target fluorescence detector. This is not the case for the Accuri C6 scatter detectors, for which especially the side scatter is less informative. We would like to highlight that the scatters have a different technical setup compared to the fluorescence detectors of the latter. The FACSVerse side scatters contain photomultiplier tubes (PMTs) for all its detectors, which can increase the signal up to 10^7 electrons per photon. Additionaly, the FACSVerse is equipped with a bandpass filter in front of the PMT, which will discriminate frequencies and denoise the incoming signal [23]. This is not the case for the

Accuri C6 scatter detectors, which contain diodes that do not enhance the signal [22]. In addition, we note that the FACSVerse instrument benefits from an improved optical bench opposed to the Accuri C6 in order to reduce the loss of signal intensity, yet resolving power based on fluorescence information was comparable.

Detector elimination and importance quantification

Our objective was to reduce the set of detectors as much as possible while retaining an optimal identification of bacterial populations. In order to do so a backward detector elimination strategy was employed (see Algorithm 1). In this way, flow cytometric data were artificially stripped, removing the least informative detector at every step of the analysis. As this strategy allowed for higher-order dependencies between detectors, it quantified the extent to which the full combination of scatter, target and secondary detectors could be used to identify bacterial cells. The detector elimination strategy was applied on 190 bacterial *in silico* communities for a species richness $S = 2, 6, 10, 14$ and 18 (Fig. 3).

It is expected to capture most important information in three detectors, i.e., two scatter detectors and one target fluorescence detector. In practice, the decline in performance started earlier than expected but only gradually; it became more substantial towards the end of the elimination scheme. In other words, a combination of the three best performing detectors resulted in a near optimal identification, but additional secondary detectors that captured cross-talk were part of the best performing subset. For the Accuri C6, at least one detector could be removed before a drop of more than 1% in performance was registered, for the FACSVerse this was at least five. This means that the reduced subset contained at most five detectors for both cytometers to optimally discriminate between bacterial populations. Fewer detectors were needed for a low S as opposed to a higher S. The FACSVerse was able to deliver a better discrimination between bacterial populations opposed to data coming from the Accuri C6 (see SI Fig. 9 for

a full overview), however, further standardization of the the experimental procedure including technical replicates is needed in order to make a conclusive comparison.

The longer a detector is retained in the elimination scheme for the identification of a bacterial population, the more important it is considered to be. Its importance could therefore be quantified by calculating its *average rank* for all *in silico* communities under consideration (Fig. 4). This allowed to inspect the set of detectors which resulted in an optimal identification. Moreover, as we have a large amount of *in silico* communities at our disposal, we could investigate whether the experimental procedure gave rise to a robust ranking of detectors or whether the importance of detectors depended on the microbial community at hand.

A general structure could be determined based on the detector ranking for both instruments. We were able to establish a general subset of detectors that allowed us to analyze a microbial community with adequate precision. The ranking varied slightly for increasing community complexity, however, and more importantly, the variability in detector-ranking dropped accordingly. This means that the ranking of detectors became more robust when the number of bacterial populations present in a community increased.

For the Accuri C6, the FL1-, FL2- and FSC-detectors could be considered as the most important ones, with FL1 being preferred for communities containing a lower amount of bacterial populations, and vice versa for the FL2-detector. This means that the performance did not deteriorate when FL4 was dropped out of the analysis; it only deteriorated marginally when SSC was dropped. It is useful to include the FSC-detector, despite the fact that its single detector performance was considerably lower than that of either a targeted or secondary fluorescence detector, which highlights the resolving power of the combination of scatter and fluorescence information.

For the FACSVerse we note that the three most important detectors were the FSC-, SSC- and FITC-detectors, which was the set of detectors to be expected. This means that the resolving

power of the scatter detectors influenced the outcome of the detector selection method considerably. In this case, both scatter detectors were placed in the top of the ranking, giving less importance to secondary detectors. Secondary detectors which measured cross-talk received an intermediary rank, although there was no order according to their estimated filter leakage or the mutual pearson correlation (see for example the PE-detector, which is not ranked in the top 5, but is the secondary detector for which most spillover was expected and measured). Detectors for which no filter leakage was expected and no mutual correlation was measured were placed last in the ranking.

Discussion

Biological and technical restrictions impact the use of FCM for microbial experiments. Multicolor approaches are difficult and therefore in many experiments limited to double staining. This means that modern instruments, as they contain more detectors than possible stains, measure more information than needed. Therefore, a considerable amount of fluorescence detectors only measure information due to cross-talk, however, knowledge is lacking concerning the resolving power of this additional information. We proposed a robust detector elimination strategy to evaluate in an objective way which detectors can be removed without loss of bacterial identification accuracy. This allowed us to characterize the importance of a detector and at the same time distinguish unexpected spectral behavior of SYBR Green I.

Summarizing our results, we can state that our microbial FCM analysis did not need all the detectors that are present on modern instruments. As expected, target fluorescence information combined with scatter information resulted in a near-optimal identification of bacterial communities. Secondary detectors gave rise to correlated information when cross-talk was measured, which could be used to boost the identification of a bacterial community. This is a known property of correlated non-redundant variables [31]. However, the improvement was limited, and the incorporation of one or two of these secondary detectors was sufficient. The effect became more prominent when the complexity of the community was increased. SYBR Green I gave rise to a much stronger signal in the red spectrum than was anticipated, which was reflected both in mutual variable correlations and the importance that is given to detectors that capture information in the red spectrum.

The importance ranking of detectors was robust in function of the composition of microbial communities, which increased for communities containing more species. Both identification performance and detector importance differed considerably for data retrieved from the two in-

struments, although the same methodology was applied. Scatter detectors of the FACSVerse resulted in a higher-resolving power than the ones of the Accuri C6. This can possibly be attributed to a different technical configuration of detectors, which differs between instruments for the scatter detectors but not for the fluorescence ones. However, further standardization of the experimental procedures is needed to be able to make this statement fully conclusive, for which technical replicates are needed instead of biological replicates. Note that the subset of detectors and detector ranking is subject to the interplay of the technical configuration of the instrument, the chemical properties of the staining in combination with the species that it is used for and the computational method that is employed.

Our method can be used to characterize the behavior of stains and the functionality of detectors in an independent and objective way. The creation of *in silico* communities, i.e. aggregating data coming from individual cultures, has proven to be effective, as the availability of species labels allows us to employ supervised machine learning methods. This approach has been used in the past to analyze the influence of various staining cocktails [32], or to analyze the influence of improved scatter information [33], albeit at a preliminary stage. As computational and technical resources have increased since then, this approach can now be fully exploited, for which our detector selection strategy is an example.

Driven by the focus on human cells [34], current instruments in FCM contain an increased number of fluorescence detectors [35], which is why modern instruments contain more lasers and detectors than necessary for microbial FCM. Our results motivate a shift in instrumental development, tailored towards specifics of microbial experiments. This shift implies the construction of instruments with fewer detectors and lasers, but of sufficient quality to detect smaller particles. These stripped instruments would reduce economical costs, which is still known to be a barrier for the field of microbiology. At the same time it will allow microbiologists to fully employ the strength of flow cytometry for their anticipated applications. This shift has initiated,

see for example [36–38], but is yet to be fully exploited. As the fields of dye chemistry, cytometry and machine learning have matured since then, we encourage a data-driven approach for future model and experimental procedure development.

Acknowledgements

We thank Susan Müller and reviewers for critical reading of the manuscript, whose comments improved the quality of the manuscript considerably. Peter Rubbens is supported by the Special Research Fund (BOFSTA2015000501) from Ghent University. Ruben Props is supported by the Special Research Fund (BOFDOC2015000601) from Ghent University and the Belgian Nuclear Research Centre (SCK•CEN). Cristina Garcia-Timermans is supported by Qindao Beibao Marine Science & Technology Co. Ltd., Qingdao West-coast economic new area, China. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI.

References

[1] HM Davey and DB Kell. Flow cytometry and cell sorting of heterogeneous microbial: The importance of single-cell analyses. *Microbiol Rev*, 60(4):641–696, 1996.

[2] J Vives-Rego, P Lebaron, and G Nebe-von Caron. Current and future applications of flow cytometry in aquatic microbiology. *FEMS Microbiol Rev*, 24(4):429–448, 2000.

[3] M Diaz, M Herrero, LA Garcia, and C Quiros. Application of flow cytometry to industrial microbial bioprocesses. *Biochem Eng J*, 48(3):385–407, 2010.

[4] MT Mee and Harris H Wang. Engineering ecosystems and synthetic ecologies. *Mol Biosyst*, 8(10):2470–2483, 2012.

[5] T Grosskopf and O S Soyer. Synthetic microbial communities. *Curr Opin Microbiol*, 18:72–77, 2014.

[6] K De Roy, M Marzorati, P Van den Abbeele, T Van de Wiele, and N Boon. Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environ Microbiol*, 16(6, SI):1472–1481, 2014.

[7] S Widder, RJ Allen, T Pfeiffer, TP Curtis, C Wiuf, WT Sloan, OX Cordero, SP Brown, B Momeni, We Shou, H Kettle, HJ Flint, AF Haas, B Laroche, J-U Kreft, PB Rainey, S Freilich, S Schuster, K Milferstedt, JR van der Meer, T Grosskopf, J Huisman, A Free, C Picioreanu, C Quince, I Klapper, S Labarthe, BF Smets, H Wang, OS Soyer, and Isaac Newton Inst Fellows. Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J*, 10(11):2557–2568, 2016.

[8] P Rubbens, R Props, N Boon, and W Waegeman. Flow cytometric single-cell identification of populations in synthetic bacterial communities. *PLoS One*, 12(1):e0169754, 2017.

[9] HH Shapiro. Microbial analysis at the single-cell level: tasks and techniques. *J Microbiol Methods*, 42(1, SI):3–16, 2000. Conference on Analysis of Microbial Cells at the Single Cell Level - Why, How, When, COMO, ITALY, MAR 25-27, 1999.

[10] S Mueller. Modes of cytometric bacterial DNA pattern: a tool for pursuing growth. *Cell Prolif*, 40(5):621–639, 2007.

[11] B Buysschaert, B Byloos, N Leys, R Van Houdt, and N Boon. Reevaluating multicolor flow cytometry to assess microbial viability. *Appl Microbiol Biotechnol*, 100(21):9037–9051, 2016.

[12] S Mueller and H Davey. Recent advances in the analysis of individual microbial cells. *Cytometry A*, 75A(2):83–85, 2009.

[13] Y Wang, F Hammes, K De Roy, W Verstraete, and N Boon. Past, present and future applications of flow cytometry in aquatic microbiology. *Trends Biotechnol*, 28(8):416–424, 2010.

[14] C. Koch and S. Müller. Personalized microbiome dynamics – cytometric fingerprints for routine diagnostics. *Molecular Aspects of Medicine*, 2017.

[15] SP Perfetto, PK Chattopadhyay, and M Roederer. Seventeen-colour flow cytometry: unravelling the immune system. *Nat Rev Immunol*, 4(8):648–655, 2004.

[16] Mario Roederer. Compensation in flow cytometry. In *Current protocols in cytometry*, pages 1.14.1–1.14.20. John Wiley & Sons, Inc., 2002.

[17] IP Sugar, J Gonzalez-Lergier, and SC Sealfon. Improved compensation in flow cytometry by multivariable optimization. *Cytometry A*, 79A(5, SI):356–360, 2011. 26th Congress of the International-Society-for-Advancement-of-Cytometry, Baltimore, MD, 2011.

[18] R Nguyen, S Perfetto, YD Mahnke, P Chattopadhyay, and M Roederer. Quantifying spillover spreading for comparing instrument performance and aiding in multicolor panel design. *Cytometry A*, 83A(3):306–315, 2013.

[19] FA Hammes and T Egli. New method for assimilable organic carbon determination using flow-cytometric enumeration and a natural microbial consortium as inoculum. *Environ Sci Technol*, 39(9):3289–3294, 2005.

[20] F Hammes and T Egli. Cytometric methods for measuring bacteria in water: advantages, pitfalls and applications. *Anal Bioanal Chem*, 397(3):1083–1095, 2010.

[21] EI Prest, F Hammes, S Kotzsch, MC van Loosdrecht, and JS Vrouwenvelder. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. *Water Res*, 47(19):7131–7142, 2013.

[22] BD Accuri™ C6 Flow Cytometer Instrument Manual. `https://www.bdbiosciences.com/documents/BD_Accuri_C6Flow_Cyto_Instrument_Manual.pdf`. Online; accessed 01-06-2017.

[23] BD FacsVerse™ Simply Brilliant. `https://www.bdbiosciences.com/documents/BD_Instruments_FACSVerse_Brochure.pdf`. Online; accessed 01-06-2017.

[24] Gianni Monaco, Hao Chen, Michael Poidinger, Jinmiao Chen, Joao Pedro de Magalhaes, and Anis Larbi. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics*, 32(16):2473–2480, AUG 15 2016.

[25] `http://m.bdbiosciences.com/us/s/spectrumviewer`. Online; accessed 03-05-2017.

[26] L Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.

[27] M Fernández-Delgado, E Cernadas, S Barro, and D Amorim. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J Mach Learn Res*, 15:3133–3181, 2014.

[28] A-L Boulesteix, S Janitza, J Kruppa, and I R Koenig. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov*, 2(6):493–507, NOV-DEC 2012.

[29] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and Edouard Duchesnay. Scikit-learn: machine learning in Python. *J Mach Learn Res*, 12:2825–2830, 2011.

[30] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009.

[31] I Guyon and A Elisseeff. An introduction to variable and feature selection. *J Mach Learn Res*, 3:1157–1182, 2003.

[32] HM Davey, A Jones, AD Shaw, and DB Kell. Variable selection and multivariate methods for the identification of microorganisms by flow cytometry. *Cytometry*, 35(2):162–168, 1999.

[33] B Rajwa, M Venkatapathi, K Ragheb, PP Banada, ED Hirleman, T Lary, and JP Robinson. Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier. *Cytometry A*, 73A(4):369–379, 2008. 24th International Congress of the International-Society-for-Analytical-Cytology, Budapest, HUNGARY, MAY 17-21, 2008.

[34] VBL Quixabeira, JC Nabout, and FM Rodrigues. Trends in genetic literature with the use of flow cytometry. *Cytometry A*, 77A(3):207–210, 2010.

[35] Y Saeys, S Van Gassen, and BN Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol*, 16(7):449–462, 2016.

[36] G Goddard, JC Martin, M Naivar, PM Goodwin, SW Graves, R Habbersett, JP Nolan, and JH Jett. Single particle high resolution spectral analysis flow cytometry. *Cytometry A*, 69A(8):842–851, 2006.

[37] JE Swalwell, F Ribalet, and EV Armbrust. SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. *Limnol Oceanogr Methods*, 9:466–477, 2011.

[38] SA Stoner, E Duggan, D Condello, A Guerrero, JR Turk, PK Narayanan, and JP Nolan. High Sensitivity Flow Cytometry of Membrane Vesicles. *Cytometry A*, 89A(2, SI):196–206, 2016.

Cytometer	Detector	Wavelength/bandwidth	Estimated filter leakage
Accuri C6	Laser: 488 nm		
	FL1	**530/30 nm**	**43.4%**
	FL2	585/40 nm	0.4%
Parameters:	FL3	670 nm LP	-
Area/Height	FSC/SSC		
	Laser: 640 nm		
	FL4	675/25 nm	-
FACSVerse	Laser: 488 nm		
	FITC	**527/32 nm**	**46.4%**
	PE	586/42 nm	0.3%
Parameters:	PerCP-Cy5.5	700/54 nm	0.3%
Area/Width/Height	PE-Cy7	783/56	1.6%
	FSC/SSC		
	Laser: 633 nm		
	APC	660/10 nm	-
	APC-Cy7	783/56 nm	-
	Laser: 405 nm		
	V450	448/45 nm	4.9%
	V500	528/45 nm	30.5%

Table 1: Detector setup of the Accuri C6 and FACSVerse; the target fluorescence detector is bolded. The estimated filter leakage is based on the BD Fluorescence Spectrum Viewer [25]. Note that this amount is not the same percentage used when applying compensation.

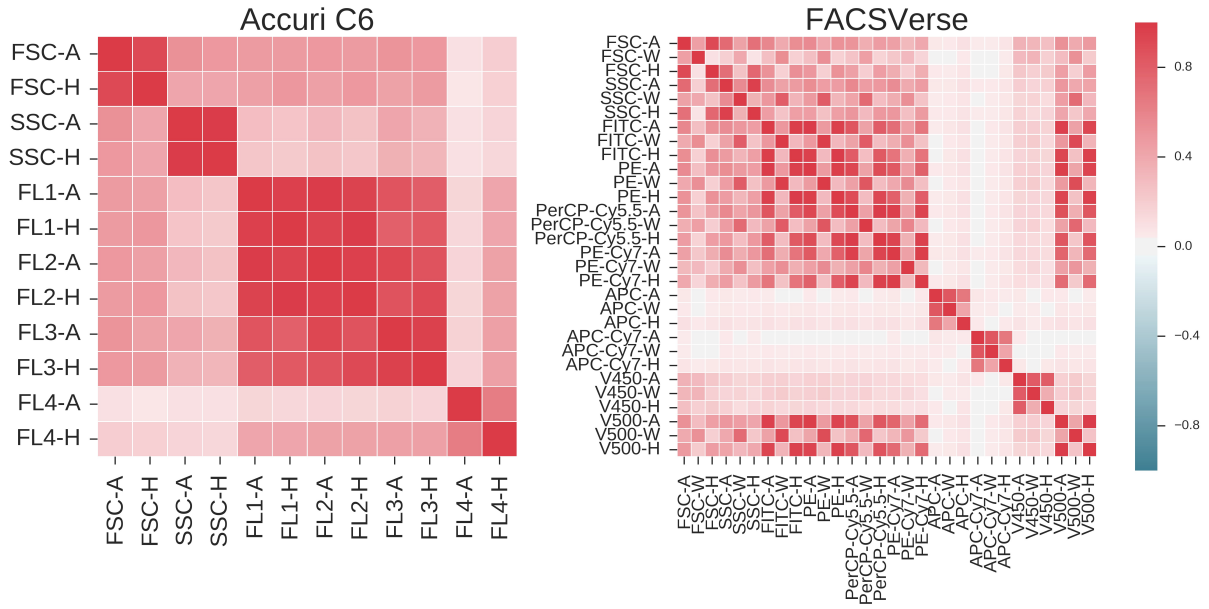


Figure 1: Average mutual pearson correlation ρ between all variables for the Accuri C6 and FACSVerse. Correlations were averaged over all individual bacterial cultures and replicate samples using a Fisher transformation; this means that ρ was calculated for $n = 40$ samples for both instruments.

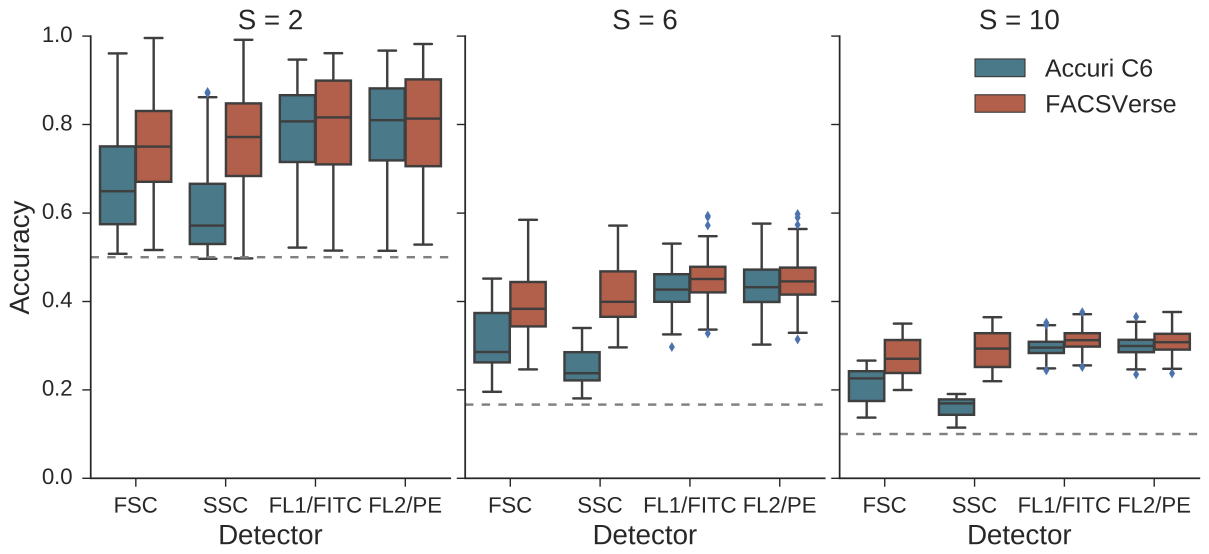


Figure 2: Single detector identification accuracies are visualized, along with the secondary detector for which the highest amount of cross-talk was expected based on the estimated filter leakage (see Tab. 1). The accuracy for a single detector was calculated for three different community sizes ($S = 2, 6, 10$), for which 190 *in silico* communities were created for both types of instruments. The box displays the 25% and 75% quartiles of the identification accuracy, while the whiskers show the full range of the accuracy, except for outliers in function of the interquartile range. The dashed line represents the identification accuracy in case of random guessing. A full overview can be found in SI Fig. 8.

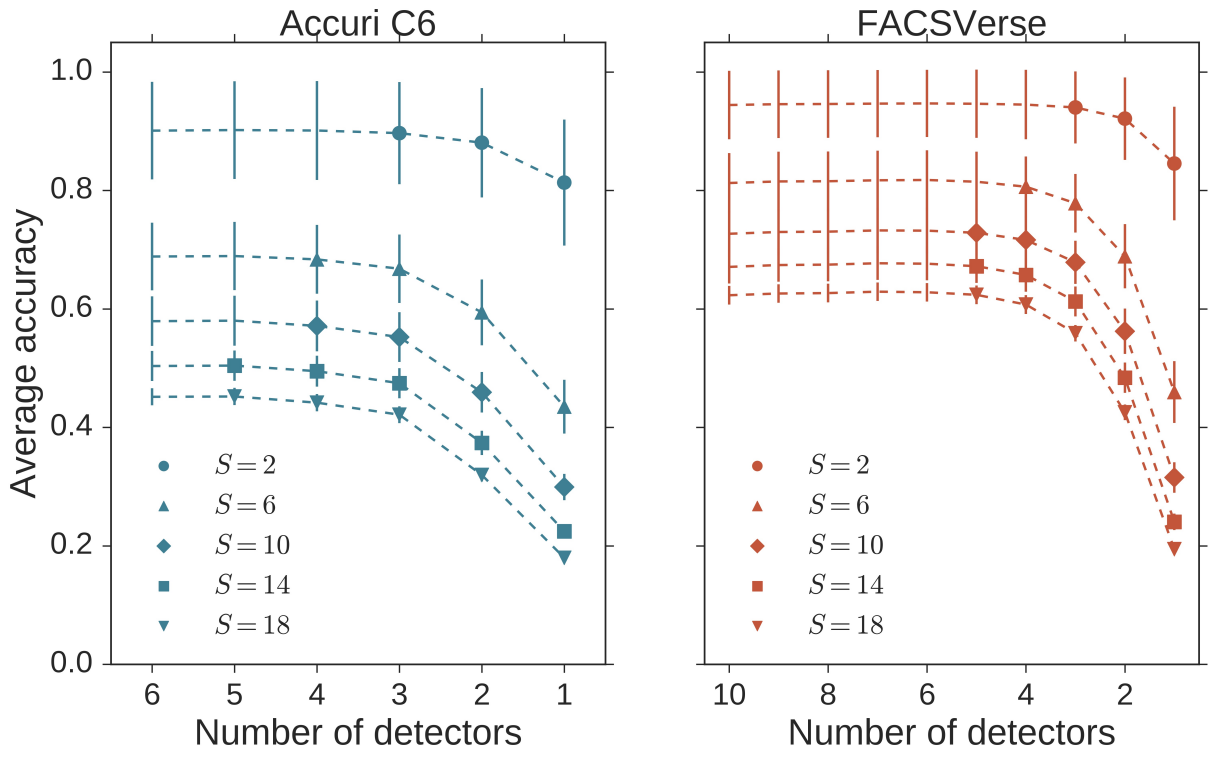


Figure 3: Average accuracies with standard deviations (SD) for 190 *in silico* communities resulting from the backward detector elimination strategy for the Accuri C6 and FACSVerse respectively. For $S = 2$ and 18, all possible community compositions were analyzed; for $S = 6, 10$ and 14, *in silico* communities were created at random, however, the same community compositions were created for both data sets. We used the Random Forest algorithm to predict the label of a bacterial single-cell, evaluated in terms of the accuracy. In order to quantify the removal of a detector, the accuracy was averaged for every S. The marker is visualized if the elimination of a certain detector resulted in a drop of more than 1% in terms of the average accuracy.

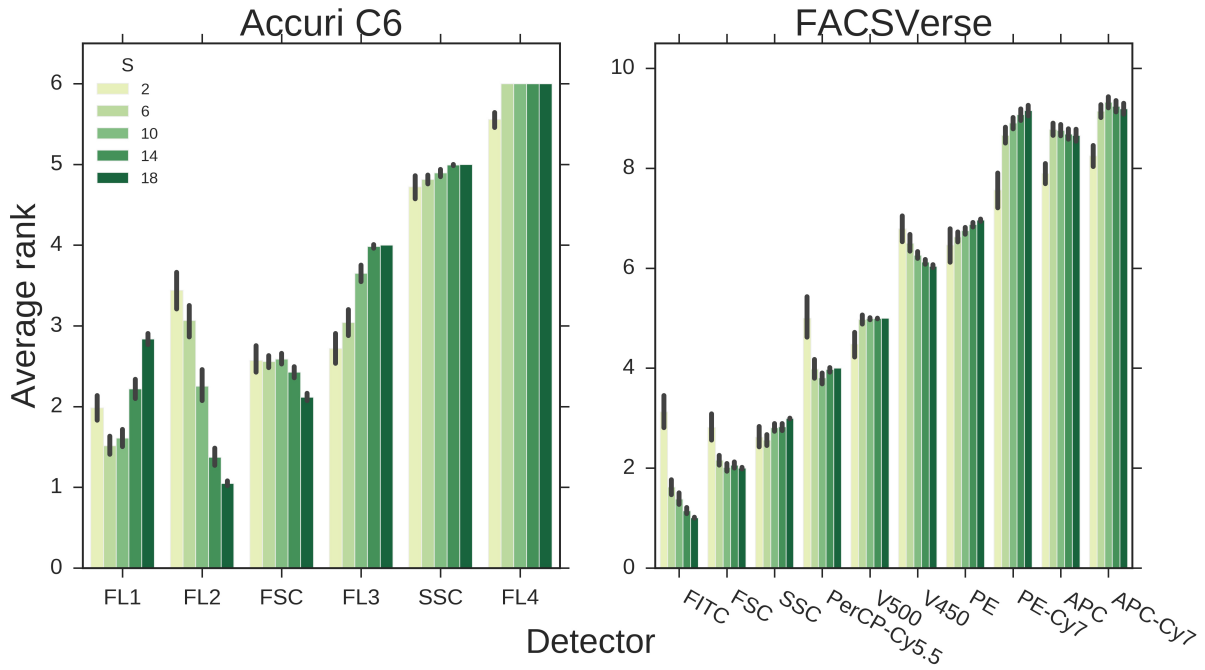


Figure 4: Quantification of the importance of detectors based on the ranking of the detector elimination strategy. To do so the average rank for a detector was determined for all *in silico* communities for varying species richness. A detector is considered important when its rank is low. Additionally, 95% confidence intervals were calculated based on 1,000 bootstrap samples. Detectors were aligned according to their total average rank, from left to right.

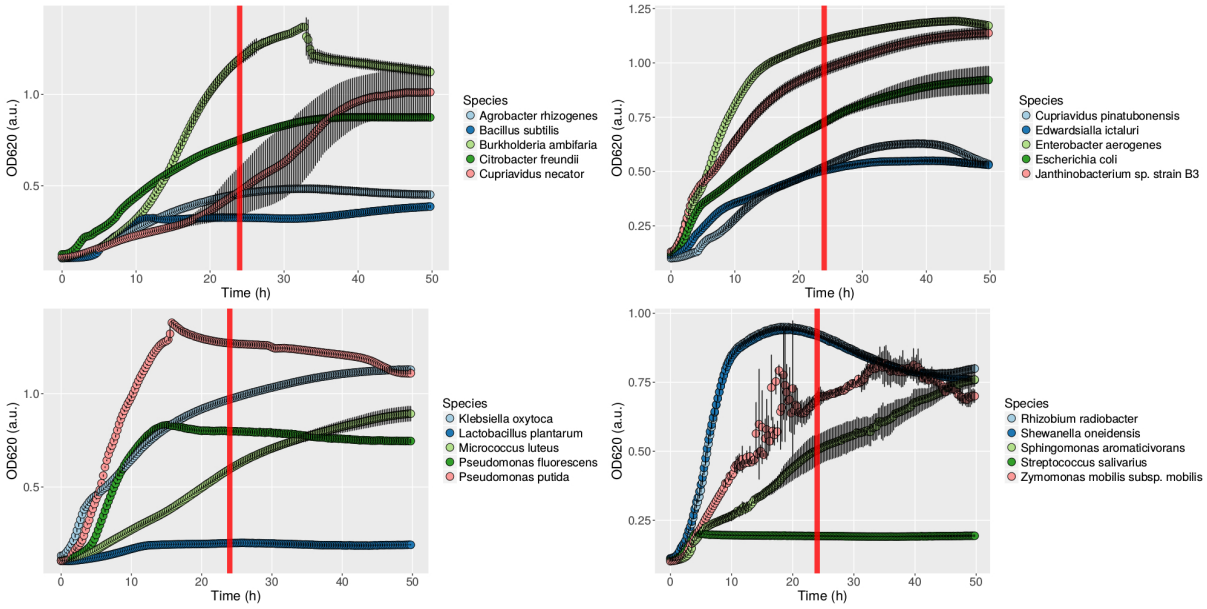


Figure 5: Growth curves based on optical density (OD) curves in function of time for all studied individual bacterial species. Cultures were sampled after 24h of incubation. It can be seen that 17 species are in early-to-mid stationary phase, 3 still were in linear or exponential growth phase.

Supplementary Information

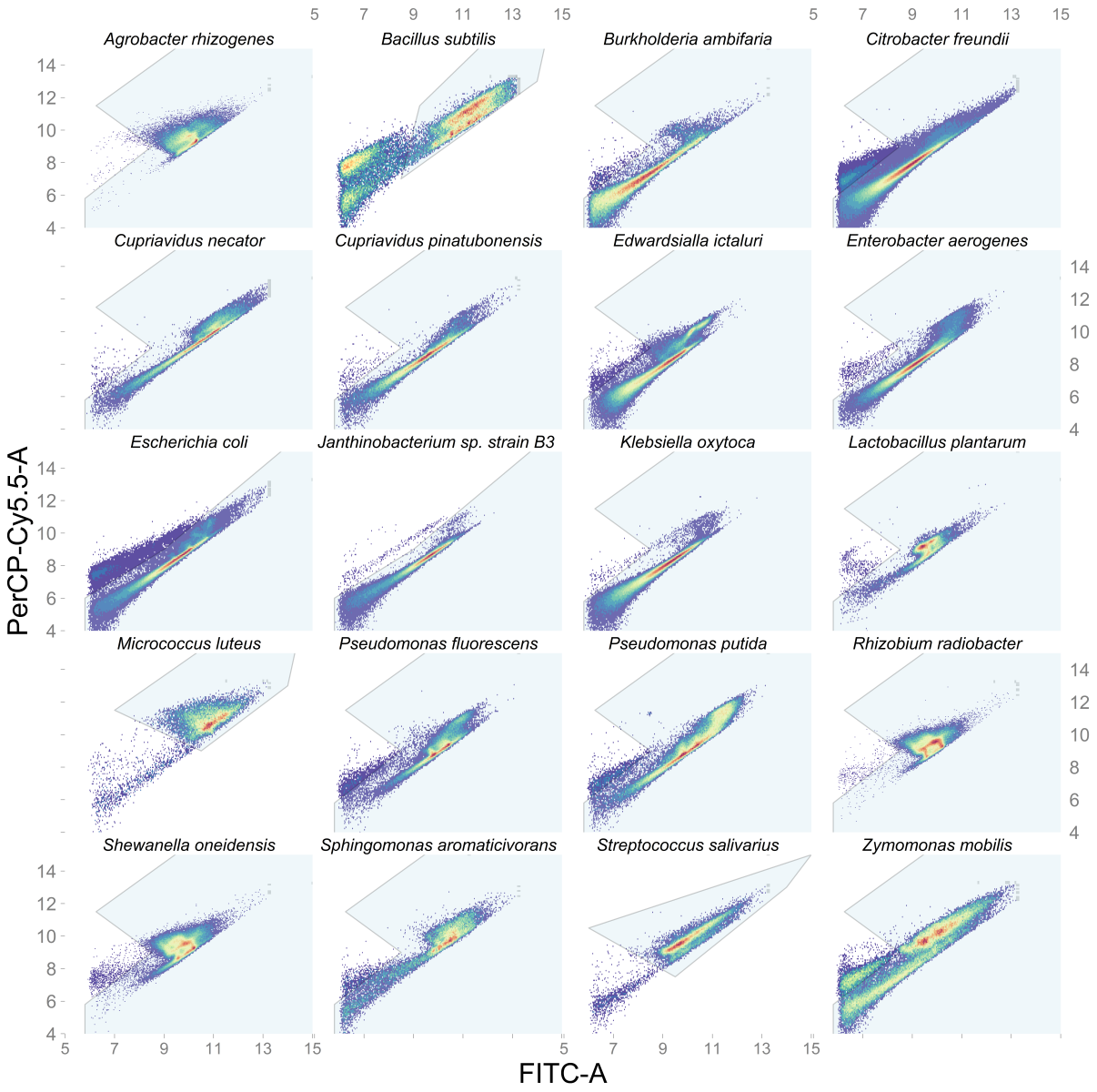


Figure 6: Visualization of the gating strategy for 20 individiual bacterial cultures using the FITC-A – PerCP-Cy5.5-A $\operatorname{arcsinh}(x)$ transformed bivariate space. Data were denoised from (in)organic noise based on a reproducible digital gating strategy and was adjusted for each culture.

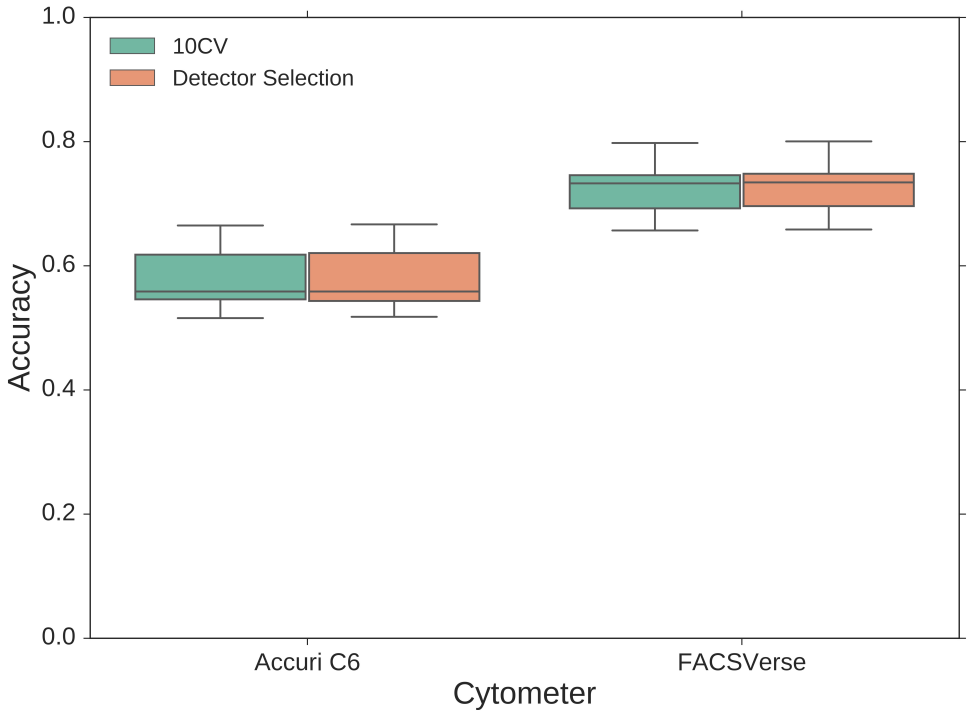
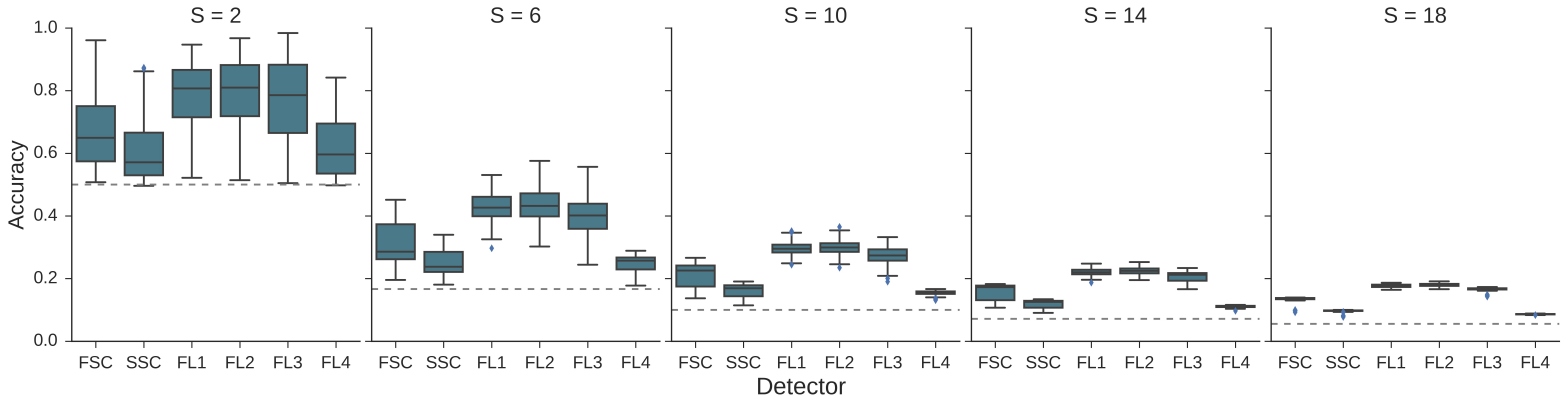
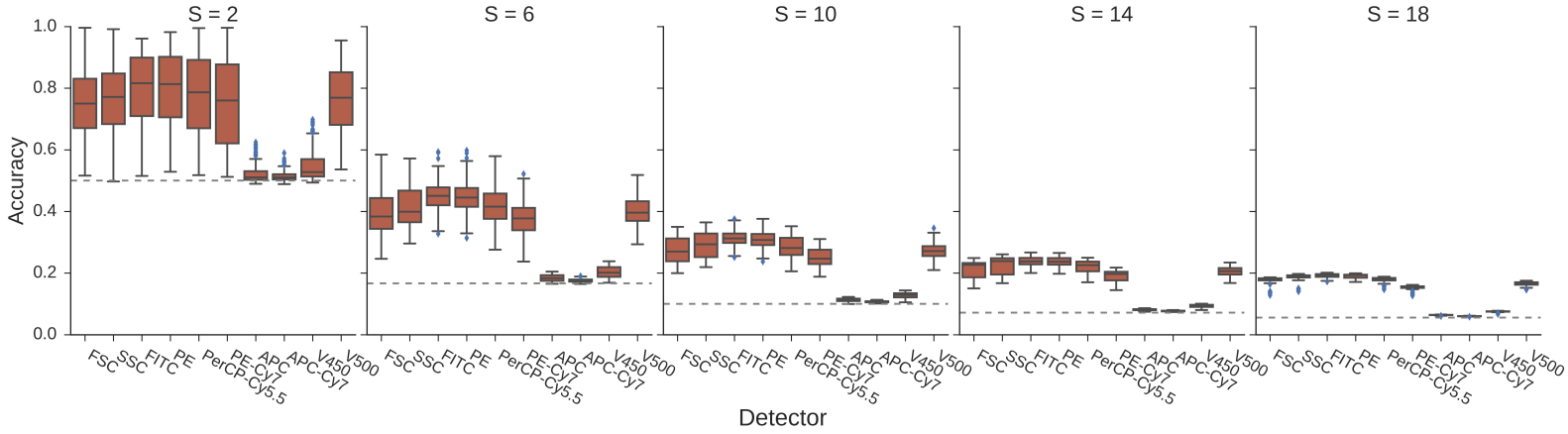


Figure 7: Comparison of twenty *in silico* communities for $S = 10$, identified with either an optimally selected amount of variables using 10-fold cross-validation (10CV) or with an optimally selected subset of detectors following a detector elimination strategy which does not incorporate cross-validation, but uses the preset $\sqrt{K}$ for the Random Forest algorithm. The box displays the 25% and 75% quartiles of the acquired accuracies, the whiskers show the full range of the acquired accuracies.



(a) Low-end cytometer.



(b) High-end cytometer.

Figure 8: Single detector performances for (a) low-end cytometer and (b) high-end cytometer. Bacterial single cells were identified feeding the information coming from a single detector to the Random Forest algorithm, for community complexities $S = 2, 6, 10, 14$ and 18; each boxplot contains 190 *in silico* communities, displaying the 25% and 75% quartiles of the dataset, while the whiskers show the full range of the dataset. Outliers are visualized in function of the interquartile range.

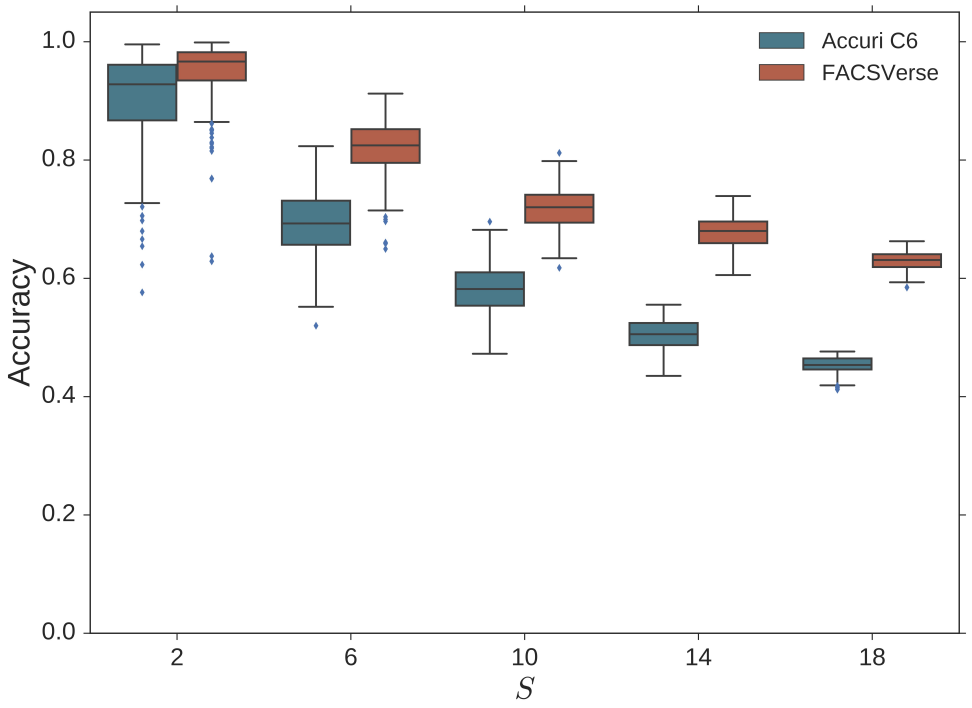


Figure 9: Resolving power in function of bacterial identification of both the Accuri C6 and FACSVerse for various S. Each boxplot contains the accuracy for 190 *in silico* communities, displaying the 25% and 75% quartiles of the dataset, while the whiskers show the full range of the dataset. Outliers are visualized using the interquartile range.